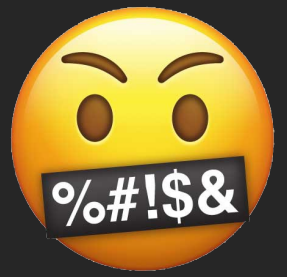
TOXIC COMMENT CLASSIFICATION

Yukhe Lavinia

- Introduction
- Toxic Comment Categories
- Dataset
- Text Analysis
- Embeddings
- Classification
- Conclusion



INTRODUCTION

Toxic comments: online remarks that are rude, disrespectful, offensive, or otherwise harmful to a discussion or its participants. They can range from mild insults to severe hate speech.

TOXIC COMMENT CATEGORIES

- **Rude/Disrespectful Language:** Name-calling, condescending remarks.
- **Obscenity:** Profanity used in an aggressive or offensive manner.
- **Threats:** Expressions of intent to harm or intimidate.
- **Insults:** Derogatory or offensive remarks targeting individuals or groups.
- **Identity Hate:** Attacks on characteristics like race, religion, gender, sexual orientation, etc.
- **Severe Toxicity:** Comments with a particularly egregious level of harmfulness.

WHY CLASSIFY TOXIC COMMENTS?

Improving Online Safety

Supporting Content Moderation at Scale

Enhancing User Experience

Enabling Safer AI/Chatbot Interaction

Monitoring and Research

Legal and Policy Compliance

Use Cases:

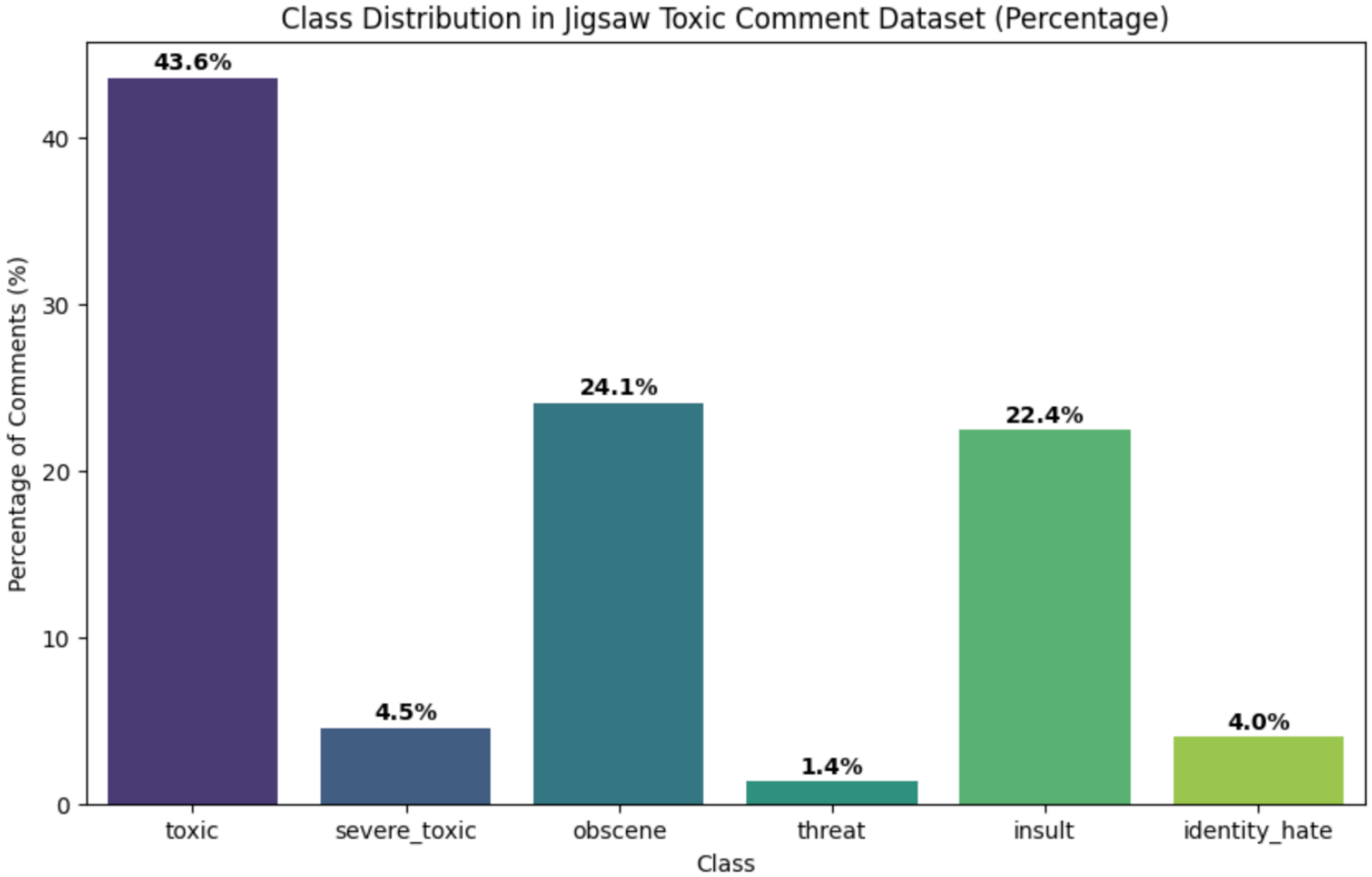
- YouTube and Reddit flagging toxic replies
- Facebook filtering hate speech
- News sites moderating comment sections
- Twitch and Discord enforcing community rules

comment_text	t
Why the edits made under my usern...	
atches this background colour I'm s...	
n, I'm really not trying to edit war. It...	
an't make any real suggestions on ...	
y hero. Any chance you remember...	

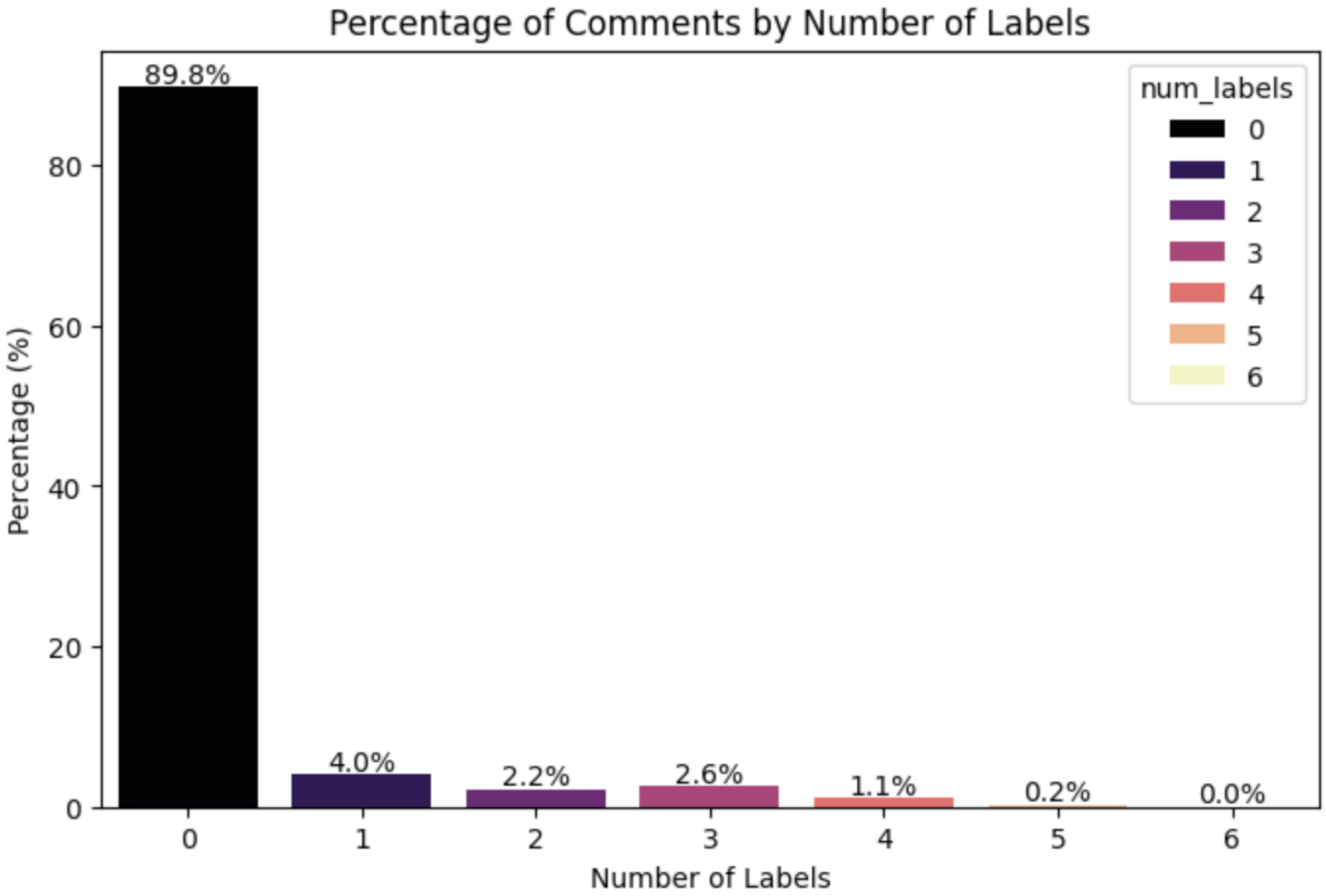
DATASET: JIGSAW TOXIC COMMENT CLASSIFICATION

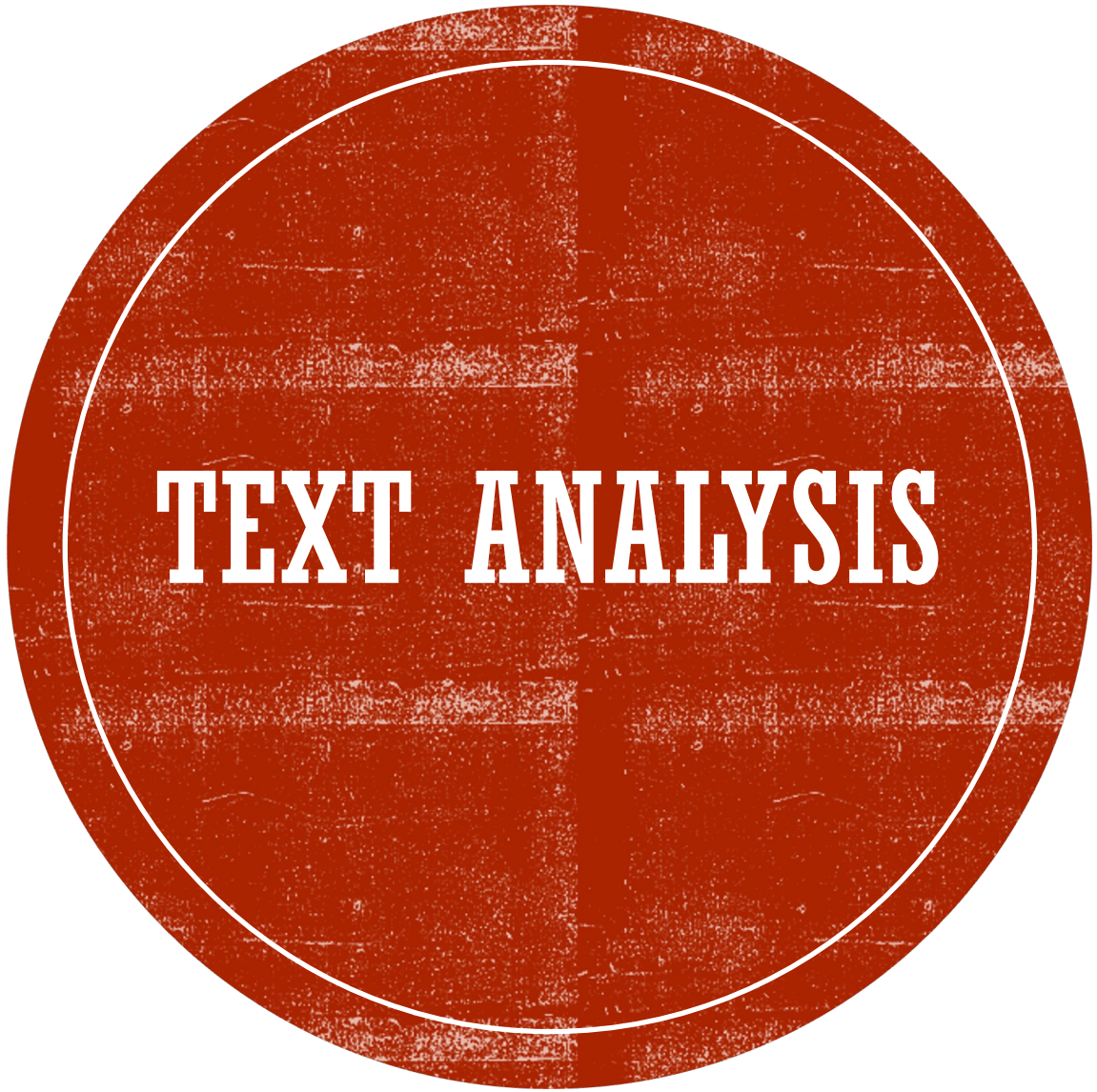
- 159,571 rows, 8 columns
- 6 classes: toxic, severe toxic, obscene, threat, insult, identity hate
- Each comment can belong to multiple classes
- Comments from Wikipedia's talk page edits

CLASS DISTRIBUTION

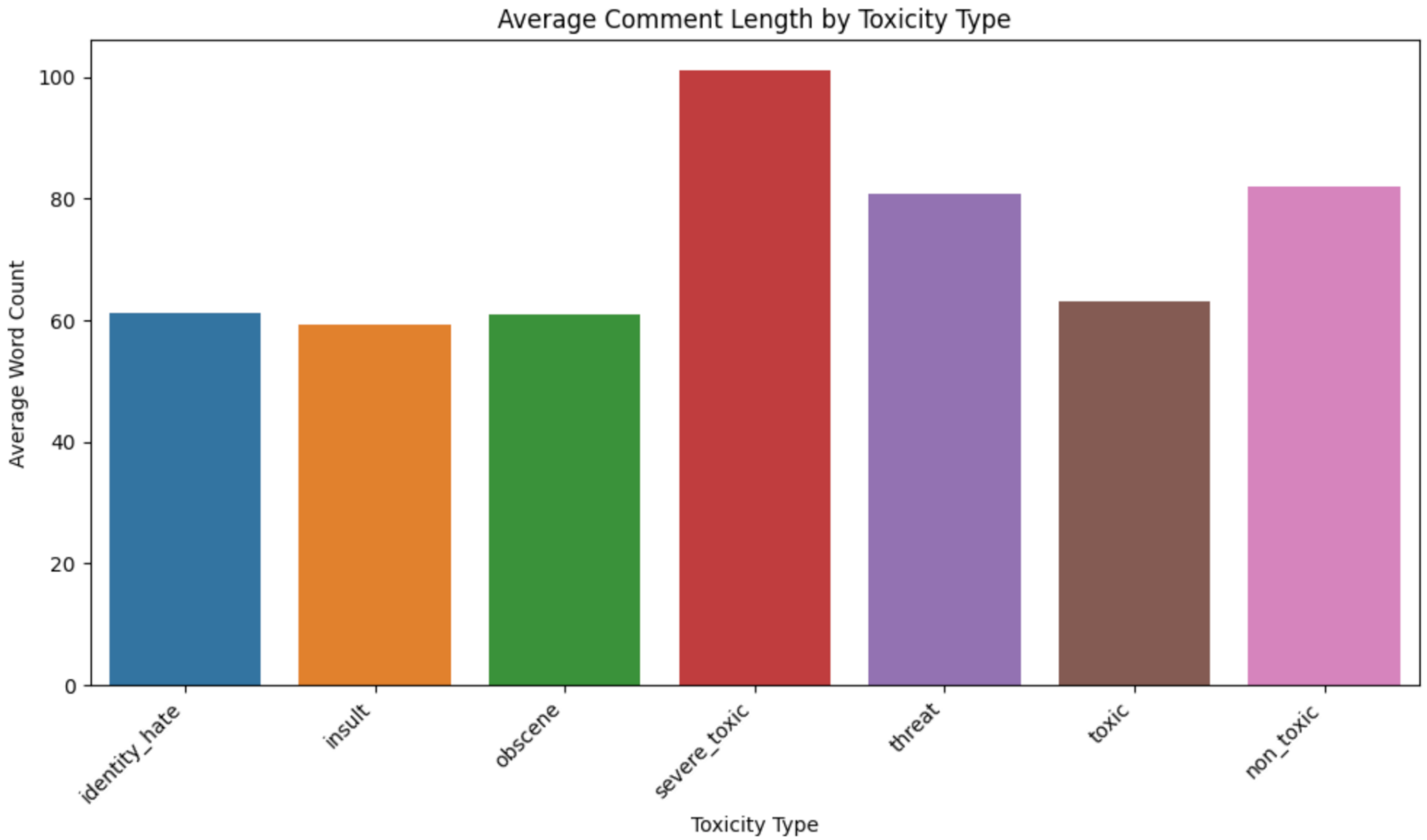


COMMENTS WITH MULTIPLE TOXICITY LEVEL

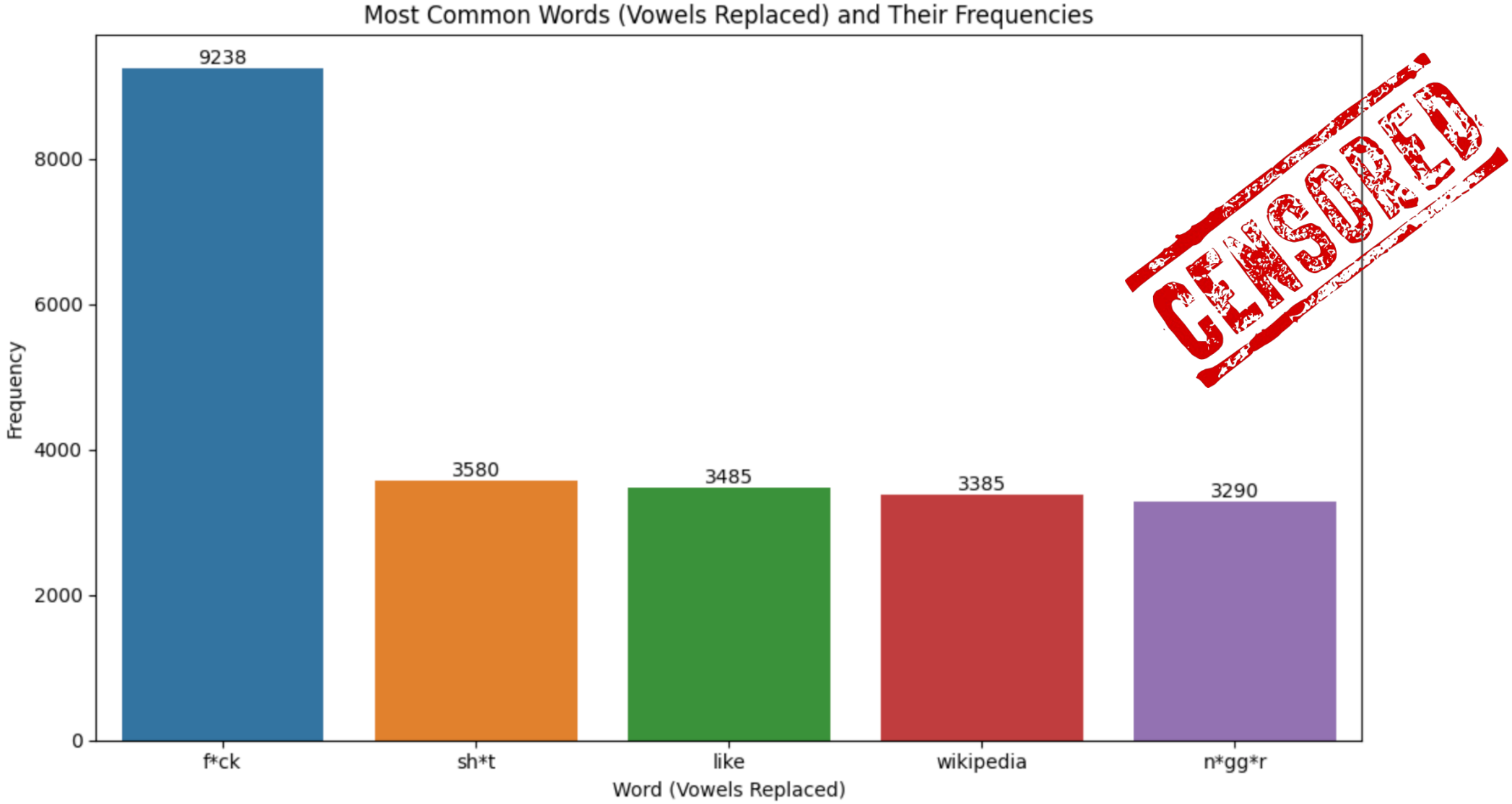



TEXT ANALYSIS

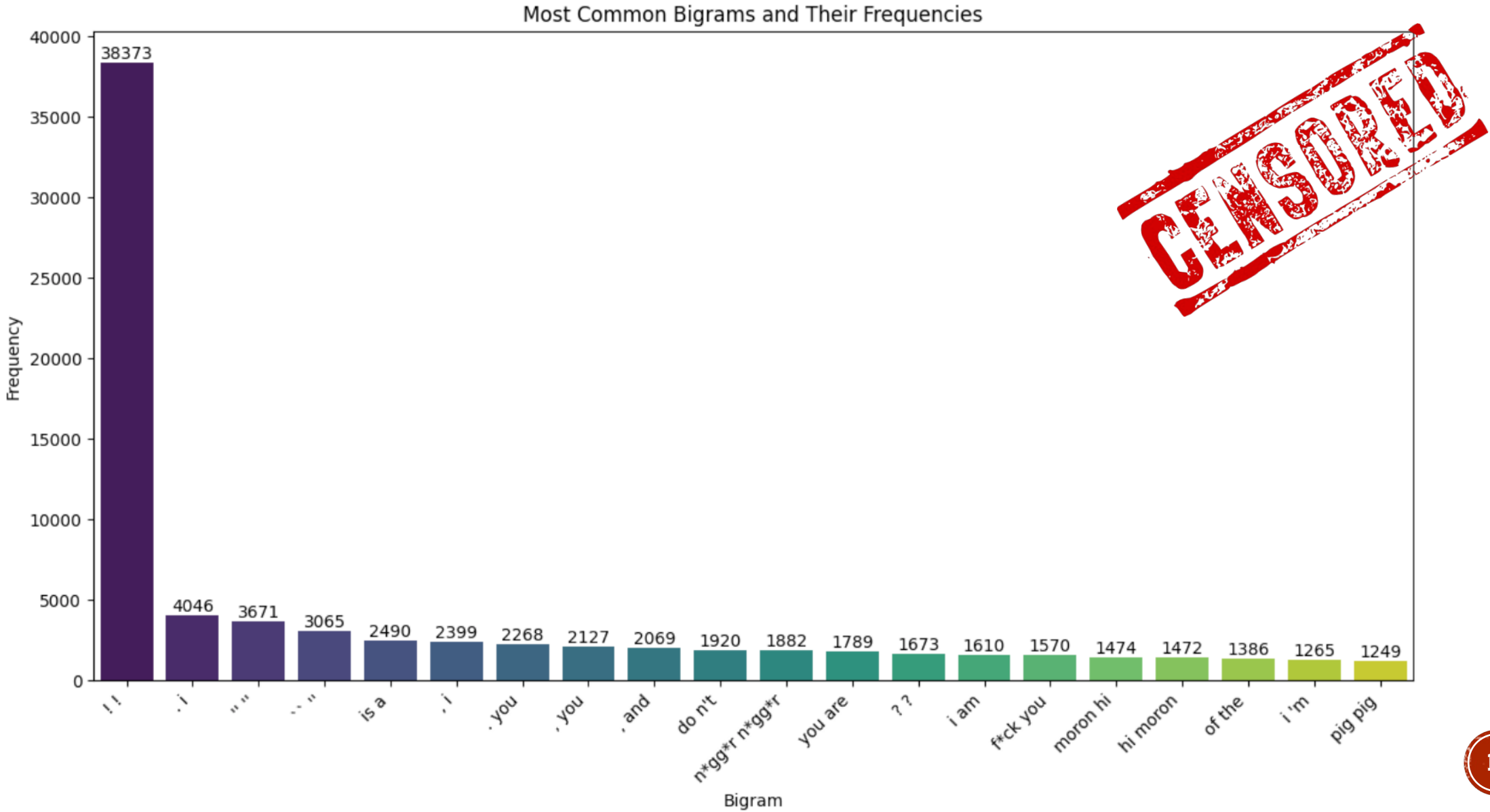
ARE TOXIC COMMENTS SHORTER/LONGER?



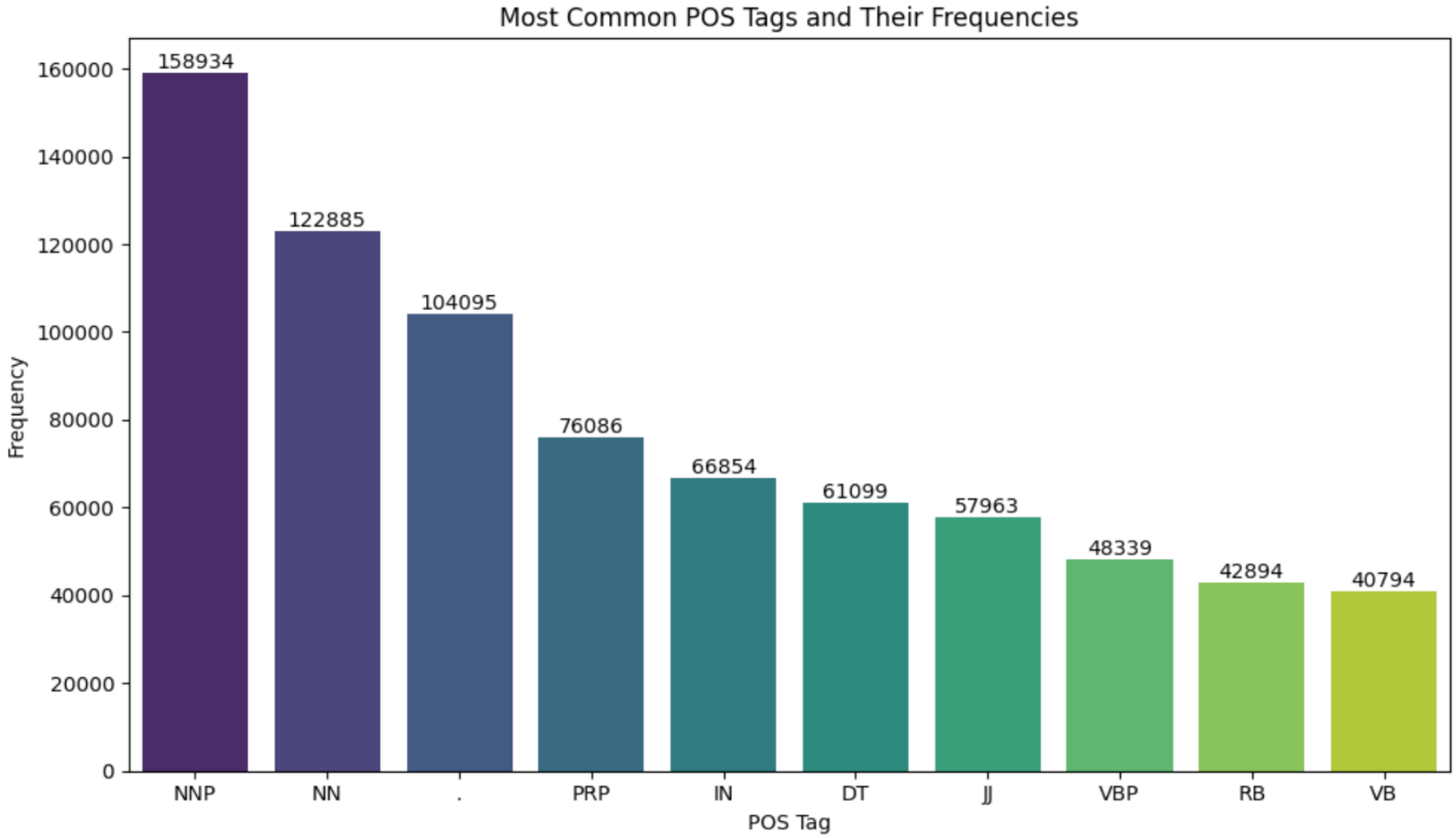
WHAT ARE THE MOST COMMON TOXIC WORDS?



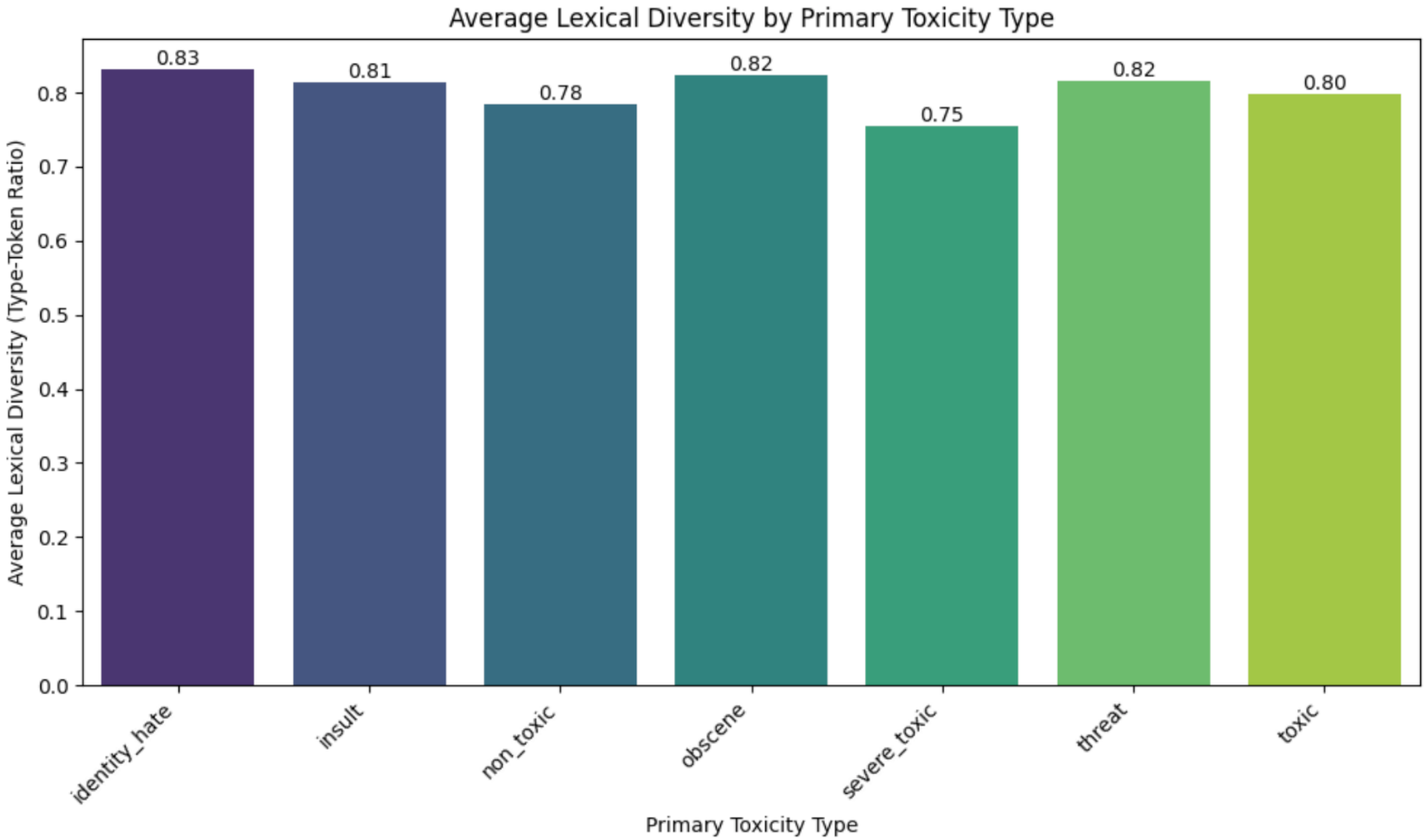
WHAT ARE THE MOST COMMON TOXIC PHRASES?

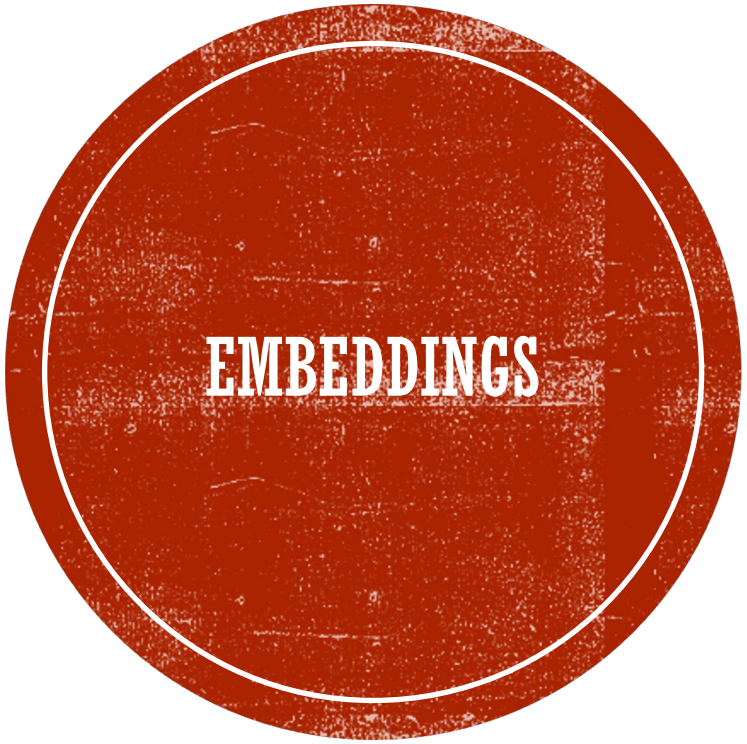


WHAT ARE THE MOST COMMON POS TAGS?



LEXICAL DIVERSITY





- Compare embeddings using GloVE and Sentence Transformer
- Visualize using UMAP
- Visualize using PCA

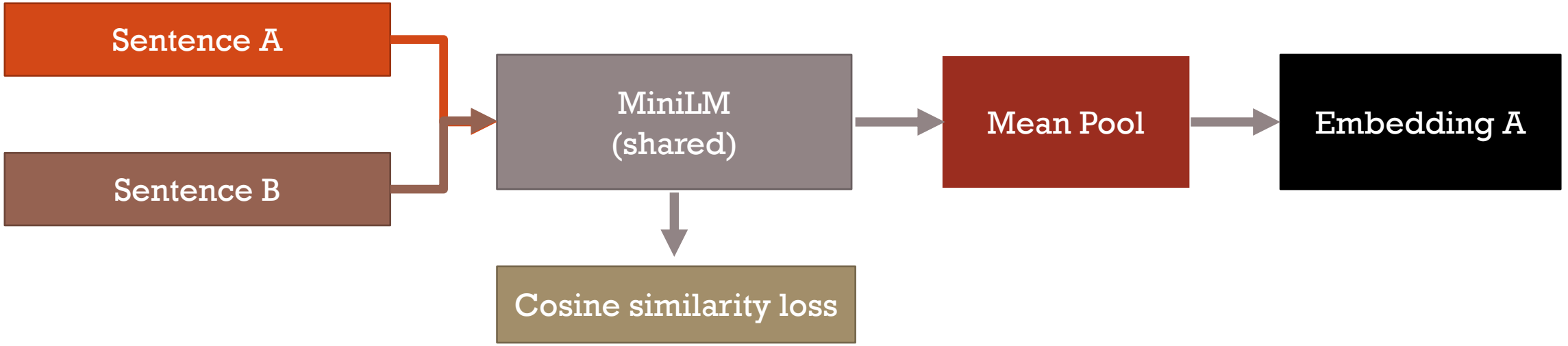
GLOVE [1]

- GloVE: unsupervised learning algorithm to obtain vector representations for words
- GloVE model is pretrained
- Training is performed on aggregated global word-word co-occurrence statistics on a corpus
- In this project the corpus is Wikipedia 2014 + Gigaword 5 (6B tokens, 400K vocab, 100d)
- Perform nearest neighbor evaluation by computing the Euclidean distance (or cosine similarity)
- Compute vector differences between two-word vectors

[1] Pennington et al., "GloVe: Global Vectors for Word Representation," EMNLP, 2014.

SENTENCE TRANSFORMER (SBERT) [2]

- A framework that modifies BERT-like models to generate semantically meaningful sentence embeddings
- MiniLM is a distilled version of BERT
- Uses Siamese-style architecture
- The encoder is shared
- Trained with sentence pairs



[2] Reimers, N., & Gurevych, I. *Sentence-BERT: Sentence embeddings using Siamese BERT-networks*. EMNLP, 2019.

SENTENCE TRANSFORMER (SBERT)

Model name	all-MiniLM-L6-v2 [3]
Architecture	MiniLM (distilled Transformer)
Layers	6 Transformer layers (L6)
Hidden size	384
Embedding size	384 (pooled output)
Training method	Sentence-BERT-style with contrastive loss
Languages	English only
Size	~80 MB

[3] Reimers, N., & Gurevych, I. *Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation*. Findings of EMNLP, 2020.

GLOVE VS. SENTENCE TRANSFORMERS

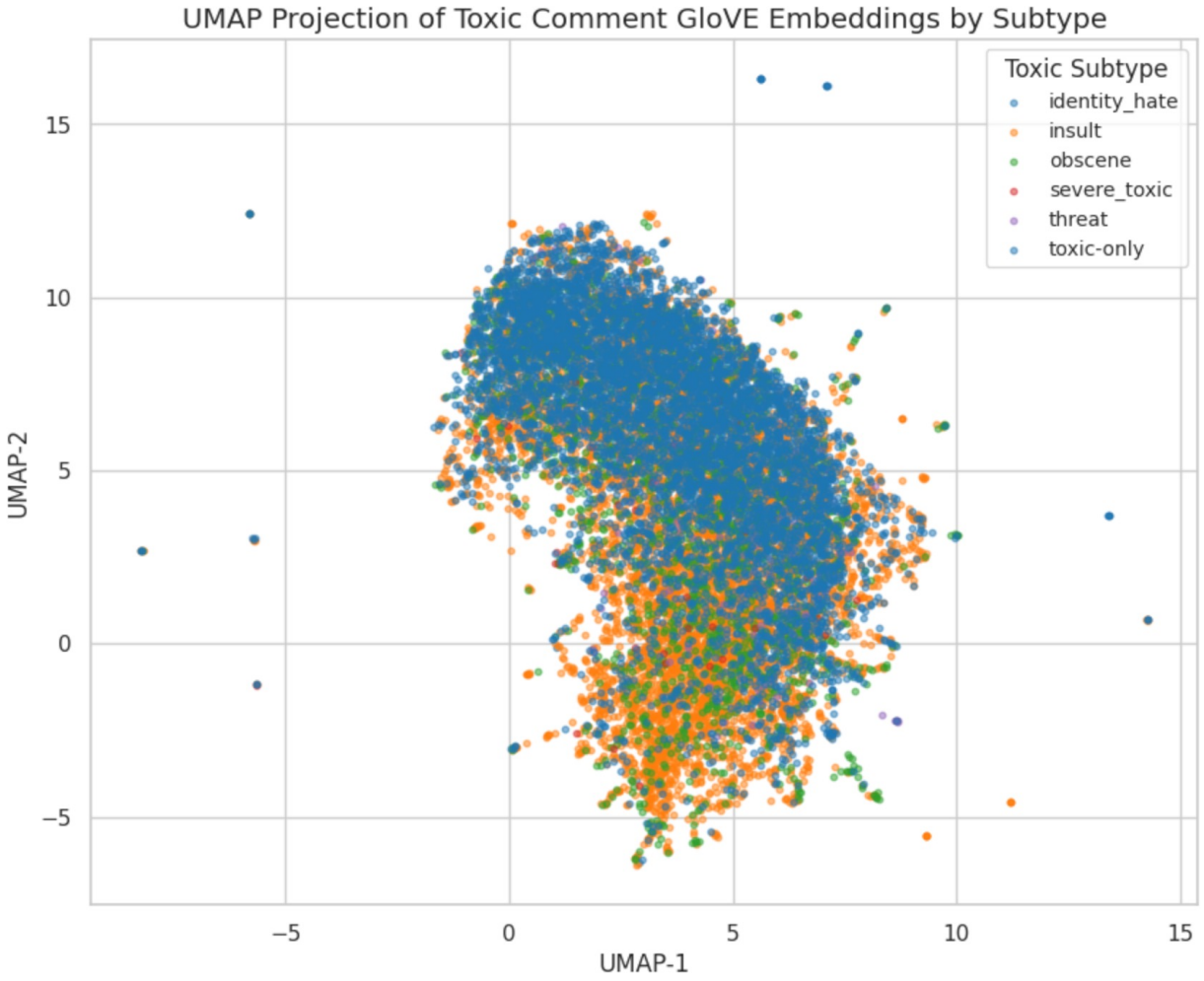
Feature	GloVe	Sentence Transformers
Embedding Level	Word-level	Sentence/Document-level
Contextual Understanding	Limited (static word embeddings)	High (contextualized sentence embeddings)
Primary Output	Word vectors	Sentence/Document vectors
Training Objective	Word co-occurrence statistics	Semantic similarity/relatedness of sentences
Architecture	Simpler (co-occurrence matrix factorization)	Complex (Transformer-based siamese/triplet networks)
Training Data	Large text corpora (unlabeled)	Semantic similarity datasets (labeled)
Main Use Cases	Word similarity, analogies, input features	Semantic search, text similarity, classification

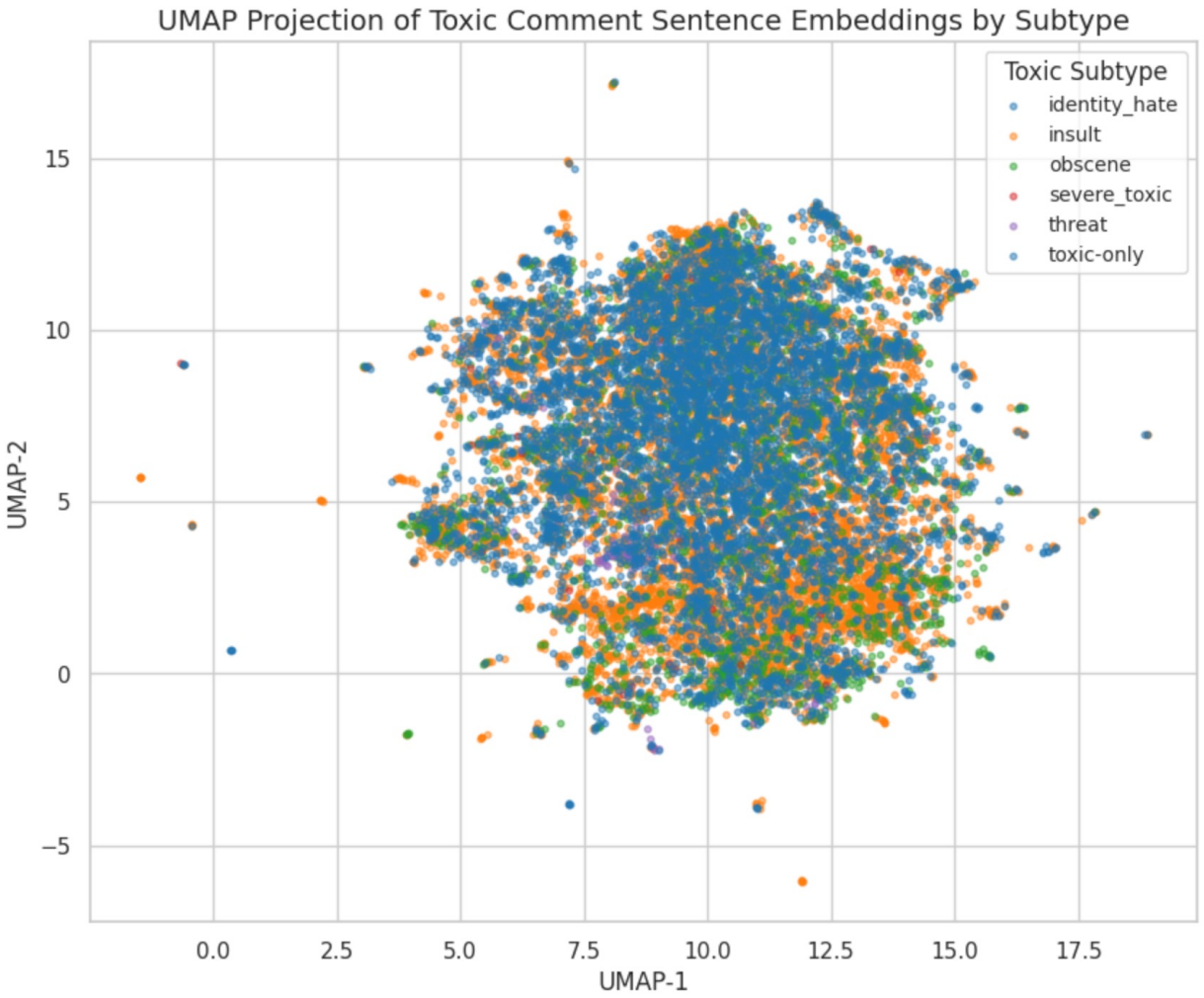
UMAP [4]

- Uniform Manifold Approximation and Projection
- Nonlinear dimensionality reduction technique
- Preserves both local and global structure
- Steps:
  - Model local relationships in high-dimensional space using k-nearest neighbors
  - Build a fuzzy graph (weighted edges between neighbors)
  - Optimize a low-dimensional layout to preserve the structure of the original graph

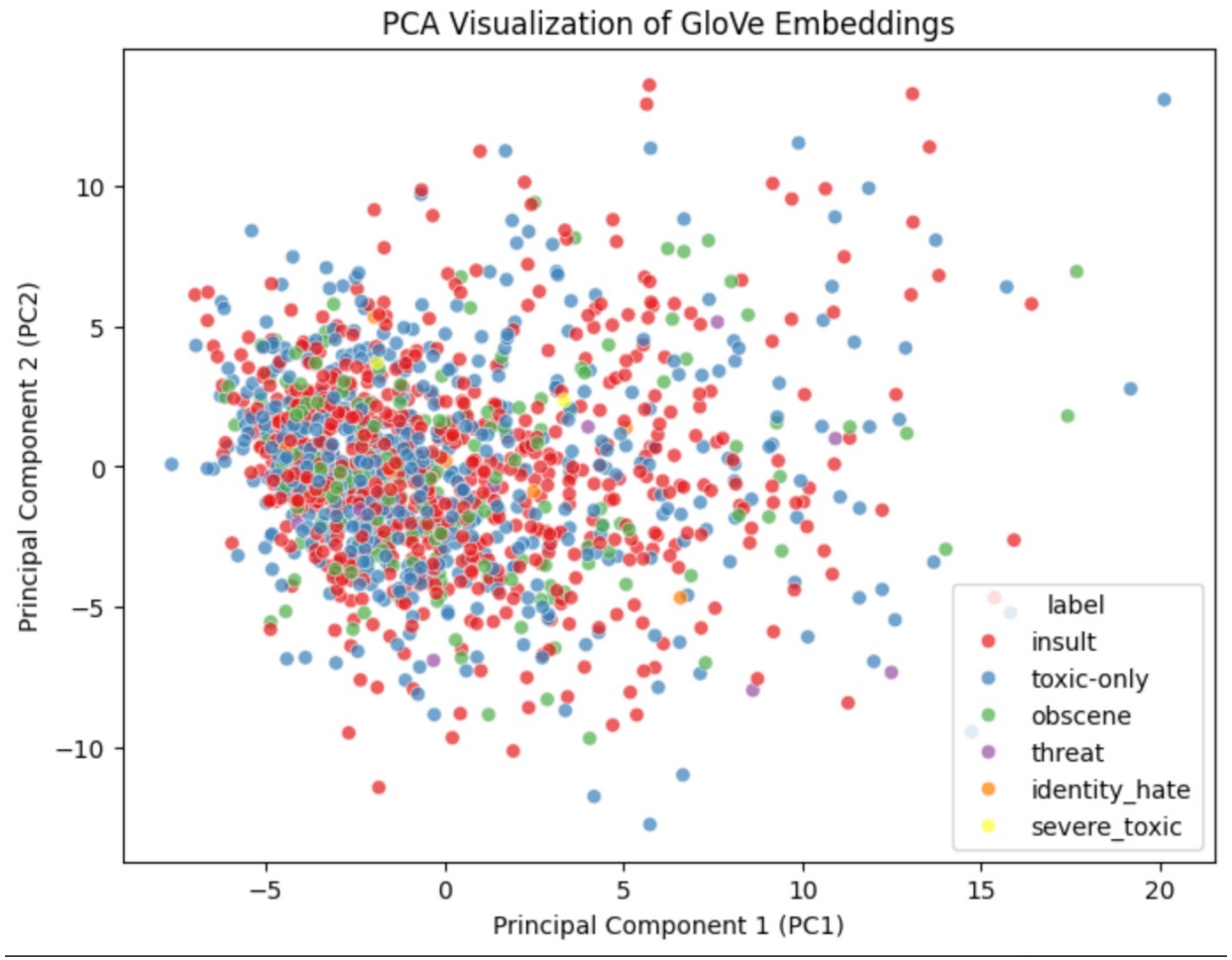
[4] McInnes et al., *UMAP: Uniform Manifold Approximation and Projection*, arXiv:1802.03426, 2018.

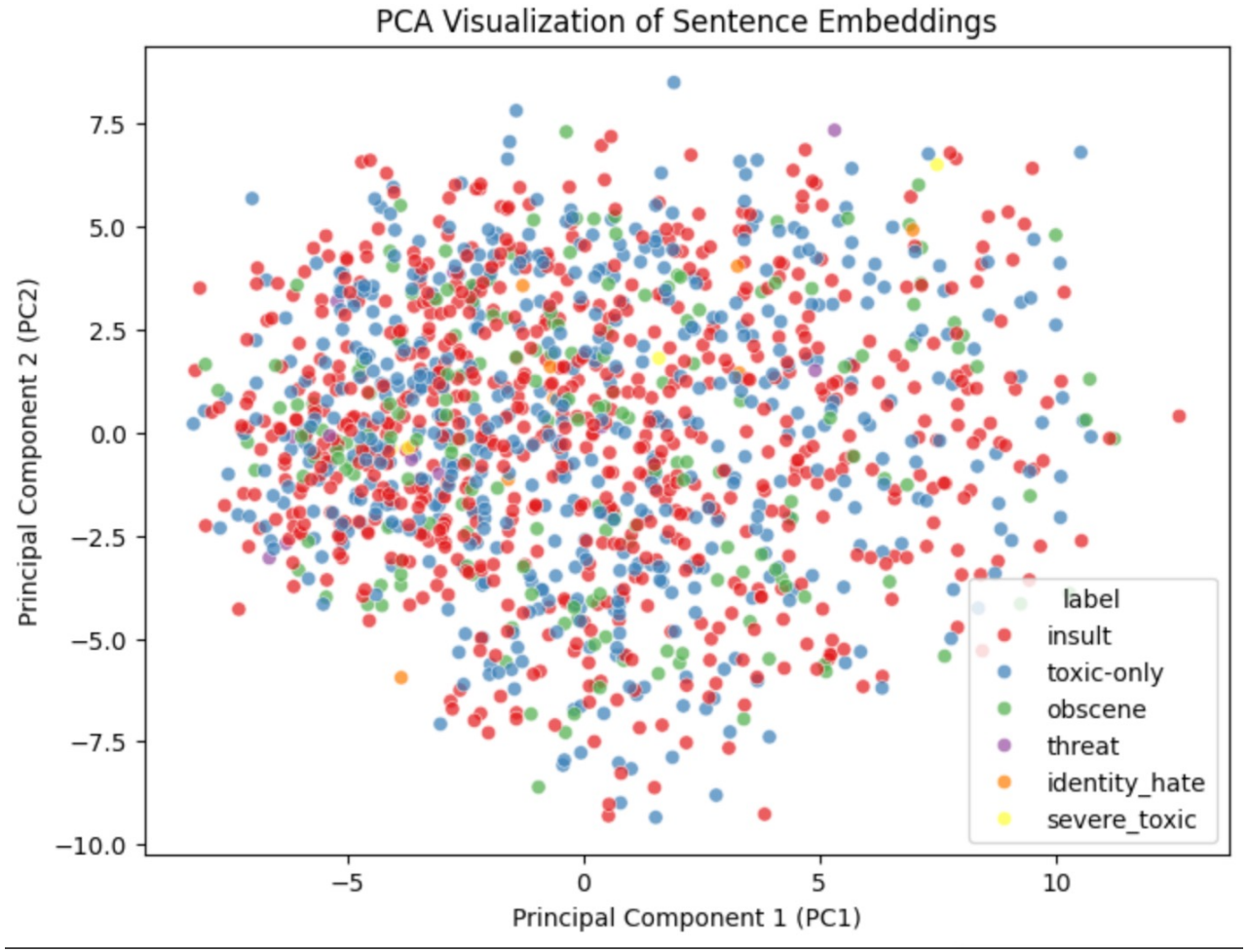
UMAP VISUALIZATION





PCA VISUALIZATION






CLASSIFICATION

CLASSIFICATION USING BERT [5]

- Used pretrained BERT ('bert-base-uncased' from HuggingFace)
- 12 layers of Transformer blocks, 12 attention heads per layer
- Hidden size: 768
- Parameters: 110 million
- Training + Hyperparameter Tuning took 1 day 56 mins 29 s
- GPU: NVIDIA GeForce RTX 4070

[5] Devlin et al., *BERT*, NAACL-HLT 2019, arXiv:1810.04805

RESULTS

Hyperparameter	Value
Learning Rate	4.04e-05
Dropout Rate	0.258
Batch Size	32
Epoch	3
Early Stopping Patience	2

Evaluation	Value
Val Loss	0.102
Val Accuracy	0.965

CONCLUSION

Text analysis

- Highest average word count: severe_toxic
- The most common toxic word is the f-word
- The most common bigram is (!, !)
- The most common POS tag is NNP
- The most lexically diverse is identity_hate

Embeddings using GloVE and SBERT

Visualization using UMAP and PCA

Classification using BERT with 0.965 accuracy

REFERENCES

- [1] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532-1543.
- [2] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. 2019 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, Hong Kong, China, Nov. 2019, pp. 3982–3992.
- [3] N. Reimers and I. Gurevych, "Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation," in *Proc. 2020 Conf. Empirical Methods in Natural Language Processing (EMNLP): Findings*, Nov. 2020, pp. 117–123.
- [4] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," *arXiv preprint arXiv:1802.03426*, 2018. [Online]. Available: https://arxiv.org/abs/1802.03426
- [5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Minneapolis, MN, USA, Jun. 2019, pp. 4171–4186.

Thank you!